Analyzing Data Science salaries and factors which define them

1. Introduction

Recent attention of public to development of AI technologies may result in bigger amount of people wanting to be part of this community as a Data Science worker. Therefore, it is important to understand what labor market rules in this field are. This paper focuses on analyzing a dataset with Data Science salaries and additional related information, such as job title, expertise level, company location and size, etc. The goal of analysis is in defining key factors which imply the salary, as well as finding other insightful patterns in data.

2. Data used

For this research, the dataset from popular data platform Kaggle [1] was used. It provides information about compensation for specialists in the field of Data Science for the years 2020 till 2023. The dataset includes information about Job Title, Employment Type, Experience Level, Expertise Level, Salary, Salary Currency, Company Location, Salary in USD, Employee Residence, Company Size, and Year. The dataset appears in .csv format and has 4976 entries.

3. Methodology

a. Overview

The data analysis was conducted in the programming language R. Code is arranged into R notebook and can be found in GitHub repository [7]. Where it was reasonable, code was arranged into functions. In other cases, these are just chunks of code.

Standard data analysis and visualization libraries were used for this research:

- maps: for visualization of maps
- ggplot2: for data visualizations
- dplyr: for data management tasks
- scales: for adding readability to visualizations

The dataset is already clean and well structured, so no additional work was required on this matter. The only fix was required to harmonize country names between dataset itself and convention used by map visualization library.

b. Linear Regression analysis

Classical implementation of Linear Regression [2] model in R was used for analyzing influence of factors on final salary amount in USD. For this, dependent variable was modeled as a sum of the following variables:

- *Year*
- *Experience Level Numeric*: here values of experience level were mapped to numerical values by the rules below. Here idea is that bigger experience level corresponds to bigger numerical value
    - Entry -> 1
    - Mid -> 2
    - Senior -> 3
    - Executive -> 4
- *Company Size Numeric*: values of company size were mapped to numerical values by the logic, similar to described above

    - Small -> 1
    - Medium -> 2
    - Large -> 3
- *Is Developed*: it is a dummy variable built based on company location column from dataset. Variable is equal to one if country at which company is located belongs to the list of developed countries, provided by United Nations report [3]. Variable is equal to 0 in other case.

Results of this analysis and its interpretation can be found in the section below.

4. Results

a. Overall description

This research started with exploration of different data distributions. The first valid one is distribution of jobs over years. As one can see on Figure 1, this number significantly increases with each next year. We conclude that the job market in field of Data Science is dynamic and fast developing. Recent attention of public to the domain of AI shows it as well.

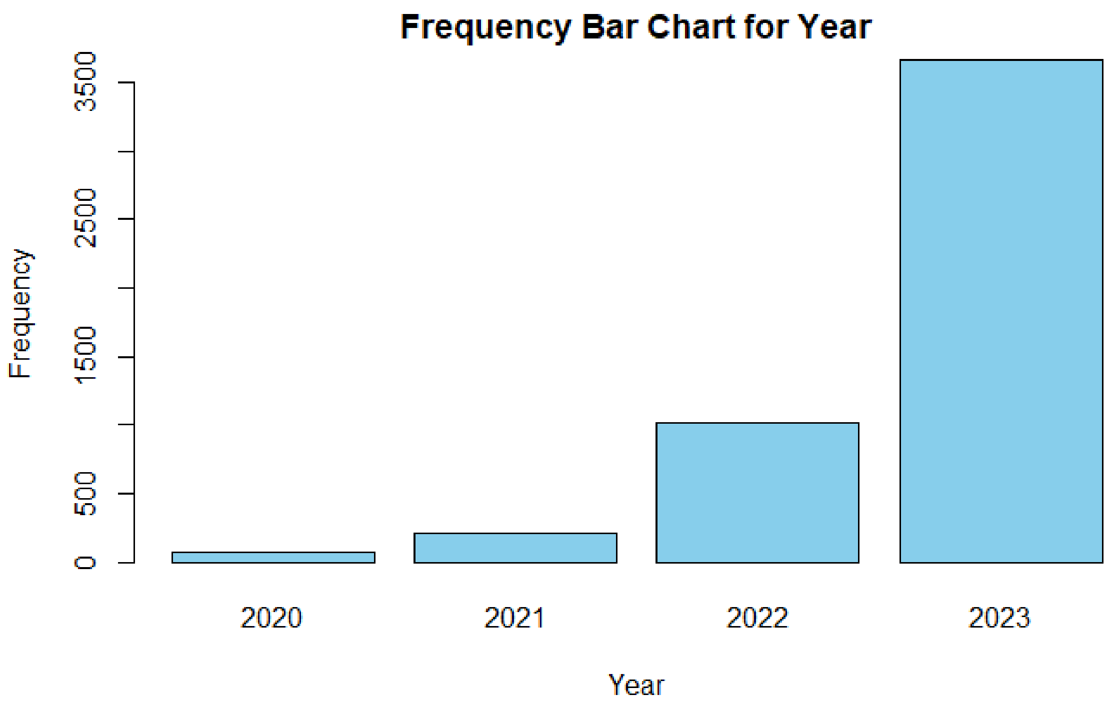

Figure 1. Distribution of jobs among years.

Other distributions, like the ones among companies, expertise level or contract type are as well of interest. While in the latter full-time jobs occupied more than 99% of the market, its visualization does not give much information. But others can be seen on Figure 2. Here we can see that most of the jobs belong to medium-sized companies, while only a few belong to large, and a very small number belong to small companies. The latter can be explained by the assumption that small companies usually don't have enough data to analyze and now enough revenues to afford a data scientist. Experience level distribution shows a bit different picture. Here more than half of jobs are occupied by senior workers, then go mid experienced, then entry and executive. The small amount of entry workers can be explained by the fact that companies are usually risk-averse, and investing in a young professional usually is risky. The small number of Executives is explained by the fact that there cannot be a lot of them because of the specifics of such expertise level. The big number of senior jobs shows that companies are more likely to work with grown professionals, and maybe pay them more, but to be sure that work will be done successfully and qualitatively.

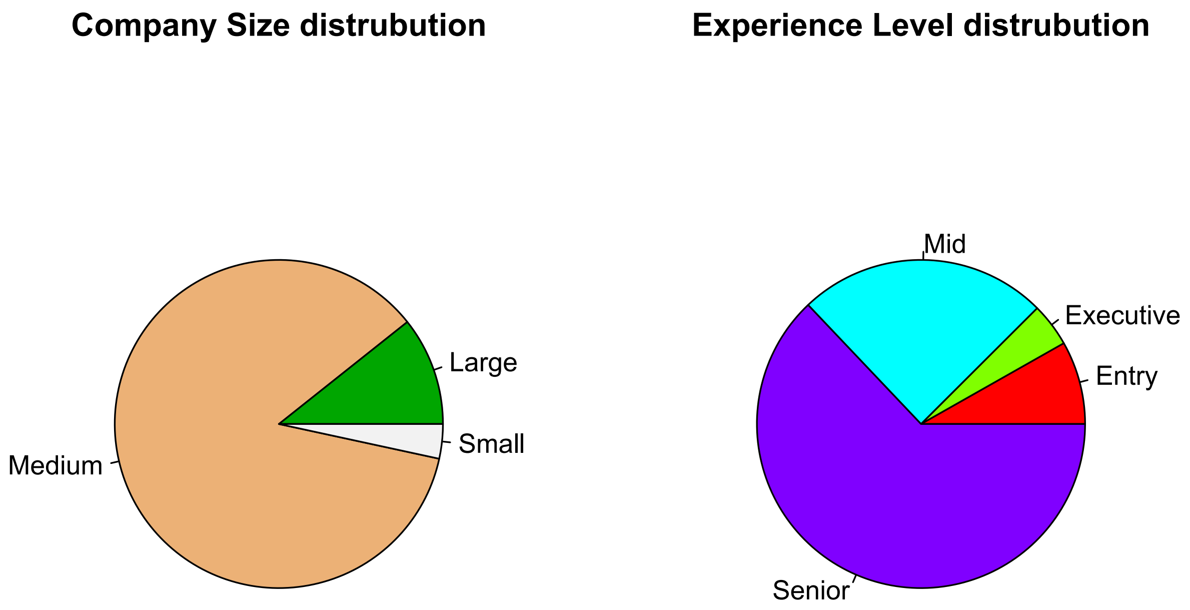

Figure 2. Company size and experience level distribution among jobs.

Another question one can raise is how related salary and other factors from the dataset are. Figure 3 shows this relation between salary and experience level, company size, employment type and year. Here are findings from this analysis:

- Experience level: as expected, with every next level salary increases. Interestingly that difference between Executive and Senior is not big, while in all other upgrades this is seen as a bigger leap. Also, with having a Mid position one has more chances to overcome higher levels by salary, as there are more outliers here.
- Company size: the biggest salaries, as well as outliers, are in medium-sized companies. Although the difference with other sizes doesn't seem to be significant. The difference is especially not significant between large and small companies.
- Employment type: Full-time job is significantly better paid than other options. The difference between the latter is not significant.
- Year: salary among years steadily increases as dates move forward. Without skyrocketing, however.

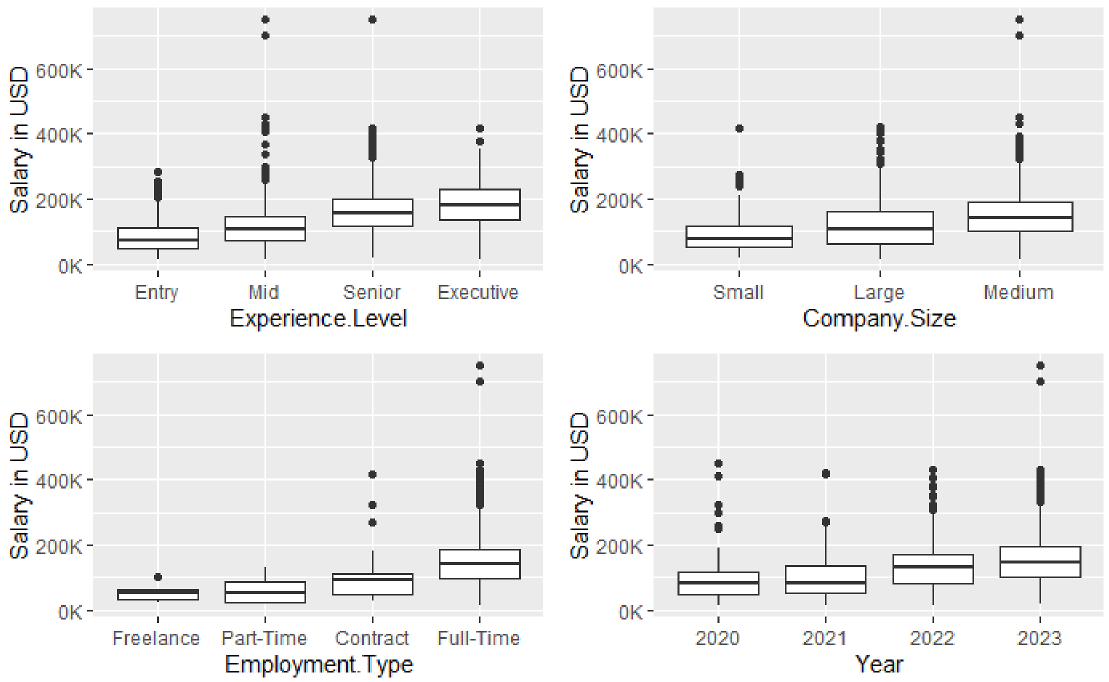

Figure 3. Boxplots of salary with regards to other factors.

b. Positions comparison

When it comes to a question which positions are best paid, one can look at Figure 4 with top median salaries. It can be easily seen that most of these positions are leader-oriented: lead, manager, director. Or one should be an architect, which requires a lot of experience and effort to achieve.

Analytics Engineering Manager	399880
Data Science Tech Lead	375000
Managing Director Data Science	300000
AWS Data Architect	258000
Cloud Data Architect	250000
AI Architect	209968
Director of Data Science	202458
Data Science Director	201000
Head of Data	200000
ML Engineer	193700

Figure 4. Best payed positions by median salary

The most popular positions among datasets are the following: Data Engineer, Data Scientist, Data Analyst, Machine Learning Engineer, Analytics Engineer, Research Scientist.

c. Countries comparison

The Figure 5 shows a visualization of a map of the world with information about median salaries in these countries. The grey fill of a country means that there are absent values for these countries. The black borders show that there are less than 5 observations for these countries. The latter was added to show that although for some countries wages may be unexpected, it may be because of lack of data available.Figure 5. Map of median salaries world. Black borders of countries mean that there are less than 5 observations for these countries.

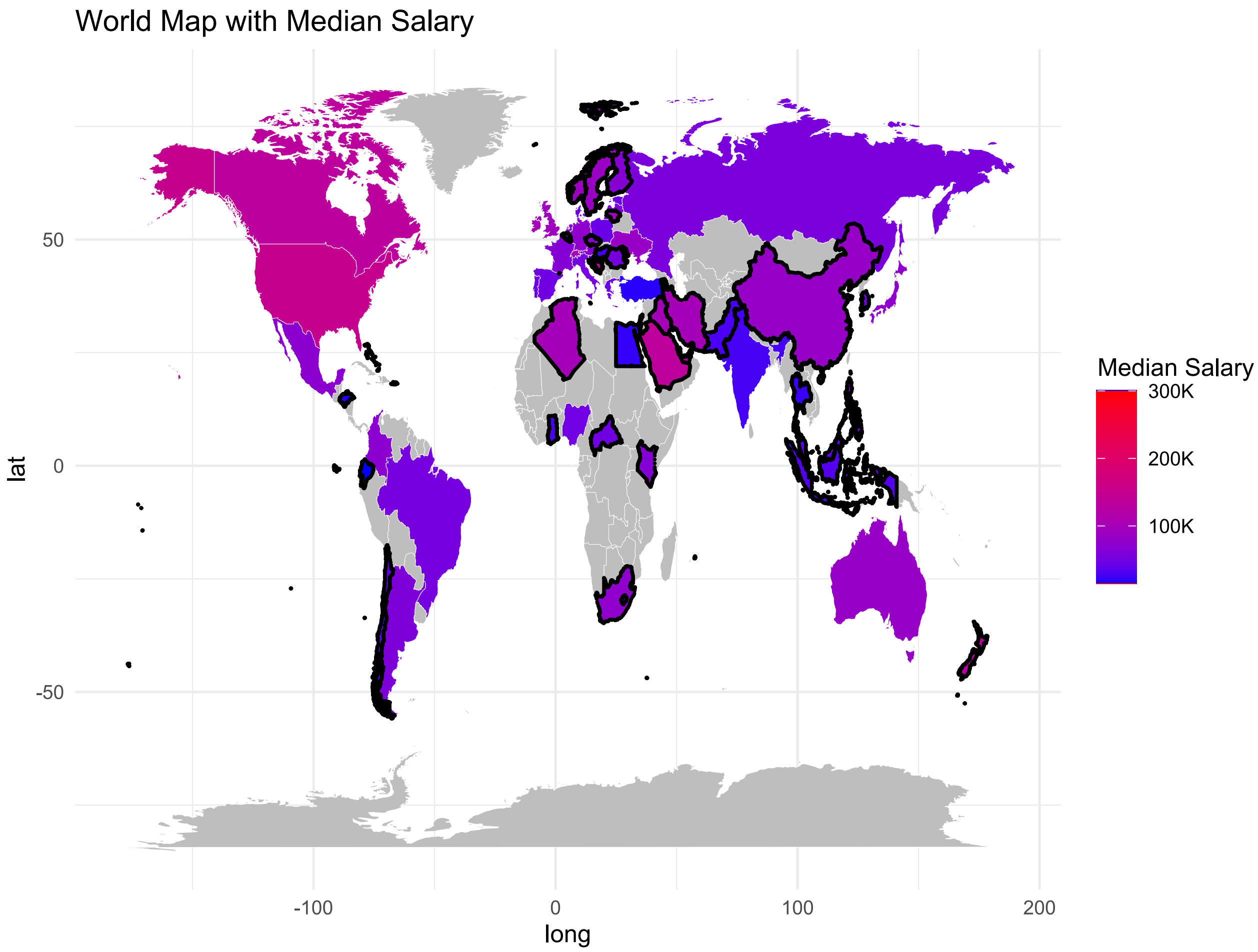

Figure 5. Map of median salaries world. Black borders of countries mean that there are less than 5 observations for these countries.

As can be seen from the picture above and from the Table 1, the highest median salaries are among the countries of Northern America and Europe. As one might expect, for most of the continents of Asia, Africa and South America salaries are low or information is not available.

Interestingly to see countries from the middle east in top as well. However, after we filter out countries with less than 5 observations, they disappear. So, although their salaries are impressive, we cannot fully rely on them, as there are not enough observations.

Table 1. Top countries by median salaries. First group includes all observations, second group is sorted to show only countries with more than 5 observations

№	Country (all)	Median salary	Country (n>5)	Median salary
1	Qatar	300 000	USA	150 000
2	Puerto Rico	167 500	Canada	132 000
3	USA	150 000	Switzerland	104 361
4	Saudi Arabia	134 999	Ireland	99 870
5	Canada	132 000	UK	92 280
6	New Zealand	125 000	Ukraine	84 000
7	Bosnia and Herzegovina	120 000	Australia	83 518
8	Israel	119 059	Japan	75 682
9	United Arab Emirates	115 000	Germany	74 597
10	Switzerland	104 361	Netherlands	73546

d. Linear regression analysis

On Figure 6 one can see the results of linear modelling, described in Methodology section. Interpretation of coefficients near independent variables shows us that size of the company does not have statistical impact on size of the salary. Year, on the other hand, shows statistical impact on the salary. Moreover, this impact is positive, which means that with each next year salaries are bigger. As one might expect, with bigger experience salaries also become bigger, and it is a statistically significant statement. As well the fact that employee works on a company in developed country significantly implies the salary. Moreover, the absolute value of this coefficient is bigger than others, from what we may conclude that it is more financially beneficial to relocate to developed countries instead of developing experience in current one.

```
Call:
lm(formula = Salary.in.USD ~ Year + Experience.Level.Numeric +
    Company.Size.Numeric + Is.Developed, data = data_df)

Residuals:
    Min      1Q  Median      3Q     Max
-145588  -43355   -8362   34281  621149

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)              -24985456    2927877  -8.534   <2e-16 ***
Year                         12347       1448   8.529   <2e-16 ***
Experience.Level.Numeric     35869       1305  27.495   <2e-16 ***
Company.Size.Numeric          2732       2436   1.121    0.262
Is.Developed                 59208       4273  13.856   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 62410 on 4971 degrees of freedom
Multiple R-squared:  0.2085,	Adjusted R-squared:  0.2079
F-statistic: 327.4 on 4 and 4971 DF,  p-value: < 2.2e-16
```

Figure 6. Results of linear regression modeling

When it comes to measuring model quality, significant F-statistics shows us that the model with these variables and coefficients explains variations with data better than Intercept solely. R-squared on level of 0.21 also shows that model explains some variations in data, although not in the best way, as good R-squared is considered to be more than 0.80. But such results of a model can be explained by the fact that not many variables were used and the information they possess cannot explain such a complex phenomenon as salary. Results may be better if we had information about revenues of companies, their domains, etc.

5. Conclusions

During this work, data about Data Science salaries and other related factors was studied. Initial descriptive data analysis was conducted, as well as more advanced analysis techniques, namely Linear Regression. For visual explanations, such visualizations techniques as map visualization, pie charts, boxplots, bar charts were used.

In summary we can conclude that the labor market for Data Science positions is healthy and dynamic: each year the number of positions is increasing rapidly, together with salaries distributions. If we measure solely my median salaries, best countries to search for a job in this sphere are in North America and Europe (USA, Canada, Switzerland, Ireland, UK). To reach the top salaries, however, specialist should develop leaderships skills together with technical ones.

References

1. SOURAV BANERJEE. (2023, December). Latest Data Science Salaries, Version 6. Retrieved December 6, 2023 from https://www.kaggle.com/datasets/iamsouravbanerjee/data-science-salaries-2023/data
2. Xin Yan and Xiao Gang Su. Linear Regression Analysis: Theory and Computing. USA: World Scientific Publishing Co., Inc., 2009. ISBN: 9789812834102.
3. 2014 WESP country classification. Retrieved December 10, 2023 from https://www.un.org/en/development/desa/policy/wesp/wesp_current/2014wesp_country_classification.pdf
4. FinalRProject, Nazar Liubas. https://github.com/liubas3171/FinalRProject